**Diabetes Detection - Group 8**

**Introduction**

The aim of the project is to analyze the dataset to detect if a person has diabetes. The project also aims to create a model that can be used to predict the chances of diabetes from individual activity, blood sample, parental heredity, food intake, sleep cycle and other factors.

**Data Description**

The dataset is publicly available on Kaggle(***Diabetes detection***). It consists of over 1 lakh rows and about 23 features.

The features include information like race, gender, age, number of lab tests performed, HbA1c test result, diagnosis, number of medications, diabetic medications, emergency visits and some question about health in past.

Here is a snapshot of the data containing some variables that we think gives better idea to detect diabetes.

Graphical user interface, table

Description automatically generated

**Table

Description automatically generated**

The response variable ‘Occurrence of Diabetes’ is categorical variable having ‘yes if diabetes detects, and ‘no’ if not detected.

**What we have done till now:**

* As a first step we pre-process the data, in which we have dealt with some missing values and outliers. Also we made two column from parental\_mother\_grandmother and parental\_father\_grandfather as these columns contains two values in each rows so we have divided this columns.
* To normalize the data set, we have created six different tables from the data set because it was not in a normal form. Diabetes\_Detection\_In\_Mother (which has data of patient’s mother and grandmother has/had diabetes) , Diabetes\_Detection\_In\_father (which has data of patient’s father and grandfather has/had diabetes), personal information(such as daily activity, age, gender, education level race), AMD\_satge(Age-related macular degeneration (AMD) is a problem with retina), reversing\_steps\_for\_diabetes ( such as taking medicines, increasing activities etc. ) , past\_health\_questions ( such as pre-diabetes detected and any other health related issues).
* Then, we have insert dataset into above tables using SQL query. And all the tables are in 3rd Normal Form.
* After that, We preprocess the data further by putting numerical values instead of categorical values where ever possible, such as gender as male, female and other takes value as 1,2,3. Also, data set has columns having values as ‘yes’ and ‘no’ where we changed that as 0 and 1.
* Moreover, to understand the data better for further processing, we visualize the data using histograms and charts. And also by this graphs and histograms, we can detect the hidden patterns which is used in model selection.

[Grishma will add this part]

**Next Steps:**

* After that, we will apply logistic regression, SVM with different kernel like polynomial, radial and linear, KNN model with different values of K, random forest and calculate accuracy and error of model and from that we will do model evaluation and after that we will predict the diabetes from that model on test dataset.
* We will also develop clusters from that we can predict that chance of occurrence of diabetes in patients from genetic features and life style.